Supplementary Guidelines, Assessment Scheme and evidence-based evaluations of the use of Evidence Based Software Engineering

Version 2.0 **DRAFT**

Austen Rainer[1] and Sarah Beecham[2]

[1]School of Computer Science
University of Hertfordshire
College Lane Campus
Hatfield
Hertfordshire, AL10 9HS
U.K.

[2]School of Information Systems
Brunel University
Computing and Mathematics
Uxbridge
Middlesex, UB8 3PH
U.K.

a.w.rainer@herts.ac.uk; s.beecham@brunel.ac.uk

University of Hertfordshire Technical Report CS-TR-xxx

February 2008

Supplementary Guidelines, Assessment Scheme and evidence-based evaluations of the use of Evidence Based Software Engineering

Austen Rainer[1] and Sarah Beecham[2]
[1]School of Computer Science
University of Hertfordshire
College Lane Campus
Hatfield
Hertfordshire, AL10 9HS
U.K.
[2]School of Information Systems
Brunel University
Computing and Mathematics
Uxbridge
Middlesex, UB8 3PH
U.K.
a.w.rainer@herts.ac.uk; s.beecham@brunel.ac.uk

University of Hertfordshire Technical Report CS-TR-xxx

February 2009

Revision history

Authors	Version	Date	Comments
Austen Rainer and Sarah Beecham	2.0	February 2009	Expansion of the Guidelines, including: 1. Empirical evidence on the use of Version 1.0 of the Guidelines 2. Abstraction of the Guidelines from a specific coursework assignment to make the Guidelines applicable to novice users of EBSE.
Austen Rainer and Sarah Beecham	1.0	February 2008	Based on guidelines provided to BSc(Hons) Computer Science students, in November 2007, as part of a final-year coursework on Evidence-Based Software Engineering.

Introduction

For the last four years of the *Empirical Evaluation in Software Engineering* final year module on the BSc(Hons) Computer Science degree programme at the University of Hertfordshire, we have been teaching and assessing students' use of the Evidence Based Software Engineering (EBSE) methodology. To aid these students in their use of the EBSE methodology, we previously developed [6] and here refine a set of supplementary guidelines for the EBSE methodology. These supplementary guidelines are intended as a kind EBSE Lite, but based on Kitchenham's original guidelines on systematic reviews [2] and Dybå et al.'s guidelines for practitioners [1]. We have also developed an assessment scheme. The assessment scheme could be used by a lecturer as part of a formal educational assessment; or by researchers, to investigate the use of the EBSE methodology;

or by the users of the methodology themselves, as a method of self-checking their application of the supplementary guidelines. Although these guidelines were originally developed for undergraduate students, we believe that they may have more general benefit for novice users of the EBSE methodology. Novice users may include software practitioners (experienced or inexperienced) who are considering using the EBSE methodology in their professional work.

In this Technical Report, we present version 2.0 of the supplementary guidelines and the assessment scheme, together with evidence on the application of the guidelines. For the assessment scheme, we do not report a specific mark breakdown for each element of the scheme. The allocation of marks to each element in the scheme is a decision that we believe that individuals should make based on their use of the scheme. For example, a lecturer may determine a particular mark allocation based on the coursework that they set (and may choose not to reveal that to the students until the assessment is complete). Conversely, a researcher may allocate marks based on the particular aspects of the EBSE methodology that they want to empirically investigate. Finally, a practising software professional may use the assessment scheme to self-check the evaluation being undertaken.

The supplementary guidelines have been developed from our experience of teaching and assessing EBSE over four years. Further information on our research work in this area is available in [4, 5, 7-11]. Our experience is that even with supplementary guidelines students often find it very challenging to undertake an EBSE exercise. Indeed, we present empirical evidence in this report of the students' opinion of EBSE. For further information on our experiences and the challenges confronting students please contact us.

Overview to Evidence Based Software Engineering

Evidence Based Software Engineering (EBSE) has recently been proposed as a methodology to help practitioners improve their technology adoption decisions given their particular circumstances [1]. In simple terms, EBSE first recommends the conduct of a Systematic Literature Review (SLR) [2, 3] to identify and appraise evidence relevant to the problem or technology under consideration. EBSE then recommends that the evaluators integrate the results of the SLR with (their) practical experience, circumstances and (professional) values. The relationship between EBSE as a whole and SLRs is summarised in Table 1.

Table 1 The relationship of EBSE and SLRS

EBSE Step	Description	Systematic Literature Review
1	Define an EBSE question to investigate	X
2	Search and select relevant literature	X
3	Appraise the selected literature	X
4	Integrate results from Step 3 with personal/practical experience and values	
5	Evaluate Steps 1 – 4 in order to improve your EBSE evaluations	

The Guidelines

1 General advice and assumptions about the evaluation

1. These Guidelines are intended to be used by 'novice users' of EBSE. Novice users may be students, researchers or software practitioners who have little experience of either technology evaluations or EBSE-based evaluations. To make these Guidelines generic, we refer to a 'task situation' i.e. a problem or situation requiring a technology evaluation. This task situation could be a professional task being undertaken by an software practitioner (e.g. evaluating one or more technologies to use in a software project or software development organisation) or an academic exercises for students (e.g. a coursework, perhaps involving a case study) or a research objective (e.g. undertaking a systematic literature review of a particular topic.)

2. For these guidelines, we assume that you as an evaluator do not have the time, effort and resources available to undertake a full-scale evaluation using Systematic Literature Reviews and Evidence Based Software Engineering. Therefore you need to think carefully about the time, effort and resources you do have available and how those resources should be allocated to the different steps of your EBSE evaluation.

3. These guidelines concentrate on performing an evaluation that typically investigates whether one thing (e.g. a tool, technology, method, methodology) is better than another thing e.g. whether the programming language C++ is better than the programming language Python. Phrased another way, the guidelines support evaluations that tend to ask questions of the form "Is x better than y?". If your evaluation is of the form "Which X's are available?" then these guidelines are unlikely to help you effectively. You may want to undertake a Scoping Study (or Systematic Mapping) instead. The structure of the particular EBSE question used in these guidelines is described in section 2. If you do not intend to develop such a structured EBSE question, these guidelines may not be particularly helpful to you.

4. You are advised to first read through the entire set of guidelines presented here, and in the Dybå et al. *IEEE Software* article (see the References section), before attempting your EBSE evaluation. It would be sensible to also consult other articles on EBSE. Some references are provided with these Guidelines. Reading through the entire set of guidelines first should better prepare you for the evaluation itself.

5. Having read through the guidelines, allocate time and effort you have available to each of the five steps of EBSE. For example, with 30 hours available, you might decide to allocate 6 hours to each of the five steps. We provide evidence later in these guidelines (e.g. sections 7 through 12) to show that some steps of EBSE are harder than others. Consequently, you may want to adjust your allocations of time and effort for each of the EBSE steps. Allocating time and effort to each step will help you prepare for the evaluation.

6. You should record your actual use of time and effort on each step as you proceed through your evaluation. Comparing your planned use of time with your actual use of time will help you, at the end of the evaluation (in EBSE step 5), to review your performance.

7. You are advised to maintain an audit trail of the degree to which you followed the EBSE guidelines. In general the audit trail can be used to help you reflect on your evaluation e.g. to

compare what you actually did against want you intended to do. Depending on the task situation, the audit trail can help you in more specific ways. For a student, the audit trail can be used to provide evidence to your assessor of the degree to which you followed the guidelines. For a researcher, the audit trail can be used as evidence in the research being undertaken. For a practitioner, the audit trail could be used to demonstrate to your peers, and perhaps to management, the rigour to which you undertook the evaluation. This could be particularly valuable when the recommendation resulting from the evaluation may be controversial to your peers. The audit trail can be used to provide evidence of the degree to which you performed each of the steps of EBSE. Evidence can take many forms e.g. screen shots of Google search results, or a copy of the first page of an article appraised. Detailed evidence could be provided in appendices to the main report required from the evaluation (e.g. a coursework, or a research report, or a internal technology evaluation report for a company).

8. You may need to refine your EBSE question as you explore the evidence available, in step 2 of EBSE, and as you start to appraise your evidence, in step 3. If you do refine your EBSE question, you are advised to state your final EBSE question in the main report and to provide the previous versions of the EBSE question in an appendix. Then, in step 5 of EBSE, you can discuss the basis for revising your EBSE question.

9. Step 4 of EBSE requires you to integrate the information identified in steps 2 and 3 with your practical experience to make a recommendation on whether or not to adopt a particular technology. Integrating the evidence can be problematic for a number of reasons. For example, you may not have much practical experience of undertaking these kinds of evaluations, or practical experience of using the technology or technologies you are evaluating. Alternatively, you may have a lot of experience of a particular technology being evaluated, and this experience may bias your judgement when undertaking steps 1 – 3 of EBSE. In sections 9 we report evidence about confirmation bias, where an evaluator concentrates on evidence that supports a particular claim (or position) and tends to ignore or devalue evidence that undermines a particular claim or position.

10. Figure 1 presents a flowchart summarising the steps of the EBSE Supplementary Guidelines.

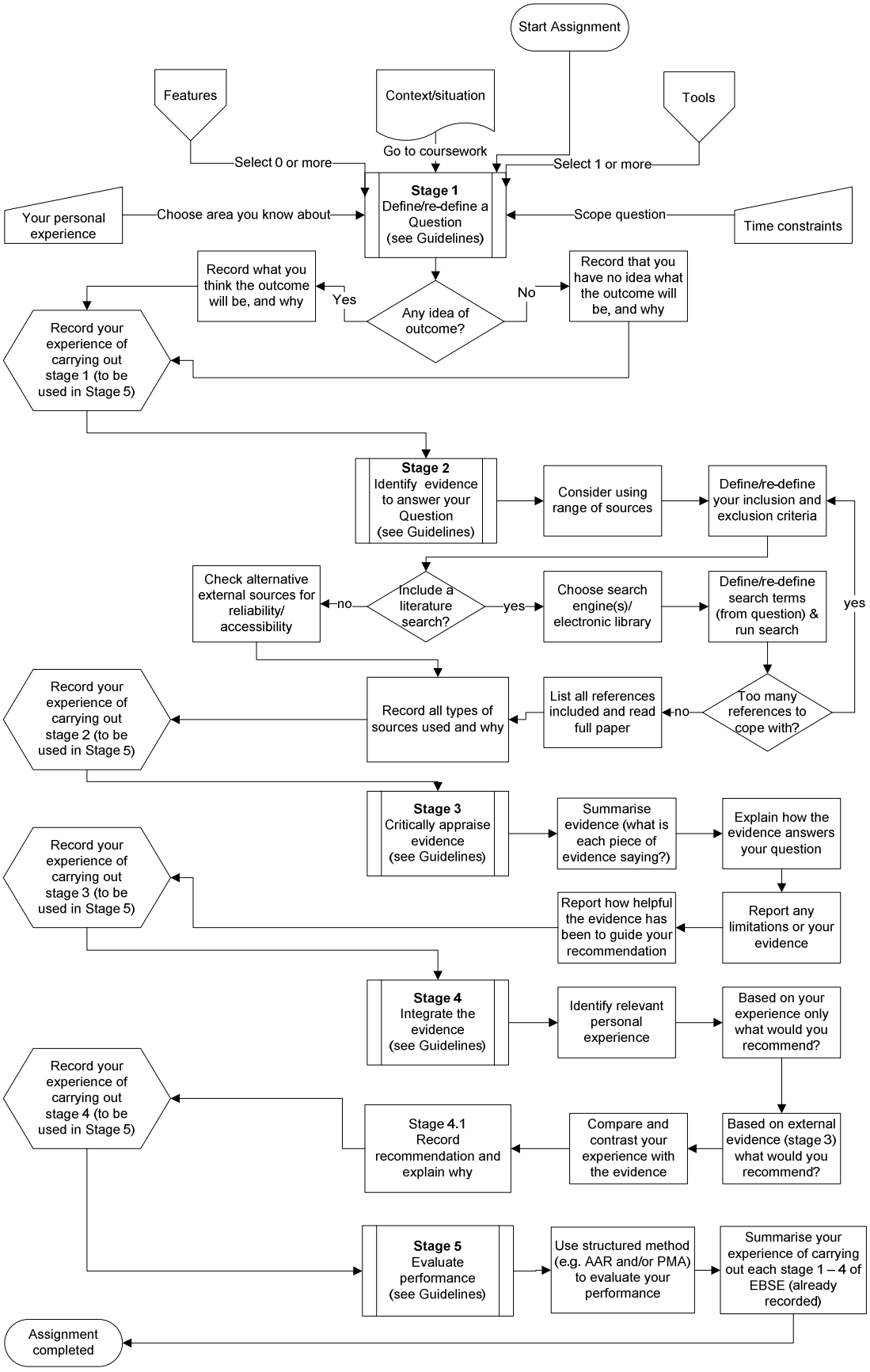

Figure 1 Flowchart of EBSE Supplementary Guidelines

2 Guidelines for EBSE Step 1

2.1 Brief description of the step

Convert a relevant problem or information need into an answerable question.

2.2 Guidelines

1. Identify problems arising in the task situation. For example, for a company evaluation there may be a range of problems relating to improving the requirements process.

2. From the problems identified, select one problem as a focus for your EBSE question.

3. State an EBSE question for the selected problem. The EBSE question should contain the following five components:

    - An intervention. This is typically a technology of some kind. In the example of the company with problems relating to the requirements process, the proposed intervention could be a requirements management tool.

    - A baseline against which the performance of the intervention would be compared. This may be another technology, or it might be not using a technology and performing a task entirely manually. In the example of the company with problems relating to the requirements process, the baseline might be a manual requirements approach (e.g. not using any tool) or a tool-assisted approach but where the tool is not requirements-specific (e.g. the company may be using Excel or Word to manage its requirements).

    - An outcome of interest. Ideally the intervention should lead to an improved outcome compared to the baseline. For the example of the company, the outcome of interest could be improving traceability between requirements (so that when one requirement changes it is easier to assess the impact of this change on other requirements).

    - A statement about the expertise and experience of the users of the technologies being compared. In the requirements management example, the company may have very experienced requirements analysts but these analysts may have no experience of using requirements management tools.

    - The situation within which the technology is being used by the users. The situation is hard to specify in detail, but would try to provide information on for example the type of software system being developed, and the size of that system. In the requirements management example, the company may develop small online databases, or may develop real-time control systems for aircraft.

4. State definitions, where appropriate, for the terms you have used in your EBSE question.

5. Explain your choice of intervention e.g. why you chose to evaluate the tool(s) that you did.

6. Explain your choice of outcome e.g. why you chose to evaluate the outcome that you did. This outcome would typically relate clearly to the problems identified in the task situation.

7. State any assumptions that you are making in your evaluation.

8. State any pre-conceived ideas you have about the expected performance of the tool you have selected for your intervention. The expected performance should be specifically related to the outcome of interest, as defined in your EBSE question.

2.3 Example(s)

One example of an EBSE question is:

> Does the use of an object-oriented programming language [the intervention]
> produce better quality code [outcome]
> compared to a structured programming language [comparison/baseline]
> when used by experienced programmers [users of the technologies]
> when developing a web-database system in a short time-frame with a small development team [situation]

Another example of an EBSE question is:

> Does the Telelogic DOORs Requirements Management Tool [intervention]
> more effectively trace requirements [outcome],
> when used by experienced requirements analysts [users]
> in the development of a large software system [situation],
> when compared to the use of no formal requirements management tool [baseline]?

In both examples, some of the terms used (e.g. experienced requirements analyst) may require formal definition and explanation. A further example is available in the "Asking the right question" box on p. 60 of the *IEEE Software* article on Evidence Based Software Engineering.

2.4 Further advice

1. Further advice on step 1 is available in the *IEEE Software* article [1] on Evidence Based Software Engineering.

2. Given the amount of detail that may need to be stated, the EBSE question may become large and potentially cumbersome. You could write a concise EBSE question, and move much of the detail to definitions and supporting explanations.

3. One of the challenges in EBSE step 1 is converting a practical problem into an answerable EBSE question.

4. Ideally, a practitioner would be interested in many outcomes for a technology of interest. You may need to adopt a 'divide and conquer' approach of separating out the outcomes into different EBSE questions. You may also want to prioritise the EBSE questions based on the more important outcomes. For a student undertaking a coursework, it would often be sensible to focus on only one outcome.

5. Consider developing several draft EBSE questions relating to the selected problem(s). The practice of developing several EBSE questions will help you to construct a better EBSE question that is relevant to the problem(s) in the task situation.

6. The definitions of terms and the stated assumptions may help you to construct search terms in step 2 of EBSE.

3 Guidelines for EBSE Step 2

3.1 Brief description of the step

Search the literature for the best available evidence to answer the question. Step 2 concentrates on the identification of evidence which is then appraised in Step 3.

3.2 Guidelines

1. Identify and state the information resources that are in principle available, those that are accessible to you (e.g. you may not have access rights to some resources) and those that can be feasibly searched with the time and effort constraints for the task situation. Item 7 in the Further Advice for this step provides some examples of kinds of information source you should search.

2. Select the information sources that you will use for your searches, state your selection, and explain your choice(s).

3. Construct and state search strings to search those information sources.

4. In addition to search strings, identify and state any other criteria you might use to either select or reject particular items of information.

5. Perform searches, providing information about the searches you conducted and summary information on the results you received from those searches.

6. Briefly review the results of searches e.g. the kinds of articles occurring in the search results. Remember that you are looking for reliable and relevant information on the outcome of a technology of interest, as defined in your EBSE question.

7. Based on your review, refine your search strings, as appropriate, and repeat searches as appropriate, in order to identify more information, more rigorous information and/or more relevant information.

8. As you perform your searches, select specific information (e.g. articles, reports, blog entries) that you will then appraise in step 3.

9. Provide a list of references to those items of information that you have selected for further appraisal. Explain your reasons for selecting those specific items of information.

3.3 Example(s)

With the following EBSE question:

> Does the Telelogic DOORs Requirements Management Tool [intervention]
> more effectively trace requirements [outcome],
> when used by experienced requirements analysts [users]
> in the development of a large software system [situation],
> when compared to the use of no formal requirements management tool [baseline]?

You might produce a table (e.g. Table 2) of information sources and indicate which are accessible.

Table 2 Example of sources of information

Source	Internet URL	Searchable
IEEE Xplore	http://ieeexplore.ieee.org	No e.g. no licence
The IEEE Computer Society Digital Library	www.computer.org/publications/dlib	No e.g. no licence
The ACM Digital Library	www.acm.org.dl	No e.g. no licence
The ISI Web of Science	www.isinet.com/products/citations/wos	No e.g. no licence
EBSCOhost Electronic Journals Service	http://ejournals.ebsco.com	No e.g. no licence
CiteSeer	http://citeseer.nj.nec.com	Yes
Google Scholar	http://scholar.google.com	Yes

For search strings, you might produce a table of sources searched, search strings used, the date of the search and the number of results. An example is given in Table 3, where the search strings are intended to find evaluations of the DOORS tool that relate to traceability. Clearly, these search strings can be refined.

Table 3 Example search strings

Source	Search string	Search date	Number of results
Google	evaluat* DOORS trace*	23 Feb 09	281,000
Google Scholar	evaluat* DOORS trace*	23 Feb 09	20,200

A further example is available in the “Asking the right question” box on p. 60 of the *IEEE Software* article on Evidence Based Software Engineering. Again, note the kinds of information reported e.g. the search string, search engine and date of search.

3.4 Further advice

1. Your searches should focus on finding information that will help you to decide whether a technology of interest (as defined in your EBSE question) positively or negatively affects the outcome of interest (as defined in your EBSE question).

2. Further advice on step 2 is available in the *IEEE Software* article [1] on Evidence Based Software Engineering.

3. There is a very large amount of public information available, much of which is searchable through various Internet search engines such as Google and IEEEXplore. This means that unless you construct sensible searches strings (and even when you do construct sensible search strings), you are likely to get a very large number of results returned from searches. You need to consider how to narrow down the searches. Some of the ways are:

   - To refine your search strings
   - To ensure your search strings focus on the intervention and the outcome of interest
   - To develop inclusion and exclusion criteria to apply to include or exclude items returned in searches e.g. to include only articles published after a certain date.

4. Even when you have narrowed down searches, you may also need to read parts of the article to decide whether it should be included or excluded. For example, you might read the abstract, executive summary, introduction and/or conclusion.

5. The *IEEE Software* article [1] concentrates on searching for research-based evidence and provides a list of helpful sources in the "Useful Information Sources" box on p. 61 of the article. The research-based evidence is likely to be more rigorous, but may not relevant. The *IEEE Software* article therefore acknowledges that you might want to consider asking (as examples) colleagues and experts, and consulting (as examples) your student notes and lecture materials, to identify relevant information. Your personal experience should be considered in Step 4 of EBSE, not here in step 2 or in step 3.

6. Declare any assumptions you make in this step.

7. Refer to the "Useful Information Sources" box on p. 61 of the *IEEE Software* article for examples of where you might search. In addition, you could also consider: blogs (e.g. Google's Blog Search or Technorati), search facilities of specific websites (such as Microsoft), search facilities for trade sites (e.g. The Register) or, if you are at a university, university-specific search sites.

8. Consider using the guidelines of systematic reviews to help you with your searches. We assume that you will not have the time however to undertake a full-scale systematic review.

9. The items that you select for appraisal in step 3 do not just have to be published, research articles. An article may be a blog entry, an email, the results of questionnaire, a web page etc.

4 Guidelines for EBSE Step 3

4.1 Brief description of the step

Critically appraise the articles identified in step 2 for their rigour and relevance.

4.2 Guidelines

1. For each article you have identified in step 2, critically appraise the article for its rigour and its relevance to the EBSE question. Use the checklist provided on p. 62 of the *IEEE Software* article to help you evaluate articles.

2. After your appraisal of an article, you may decide to reject the article entirely because the arguments and evidence in the article are either not rigorous, not relevant, or both. You should state whether you have rejected an article and then exclude that article from the remainder of the EBSE evaluation.

3. After you have appraised each article you need to consider the overall conclusion of all of the articles. This requires that you integrate the evidence and arguments from the articles in some way.

4. On the basis of the articles and information that you have reviewed in step 3 only, state a tentative recommendation about whether one should or should not adopt the tool(s) that you are evaluating. It may be that you modify your recommendation after undertaking step 4 but it is sensible to state a recommendation now.

4.3 Example(s)

It is difficult to present a detailed example here. Table 4 provides an example where 10 papers identified in EBSE Step 2 have been appraised using the EBSE checklist provided in the *IEEE Software* article [1]. The '**' notation identifies those papers that were rejected during the appraisal. Note that the table shows each paper being individually appraised. The results from the (remaining) papers would then need to be integrated.

Table 4 Example appraisal of papers using EBSE checklist

Paper	1. Vested interest	2. Valid Evidence	3 Important evidence	4. Used in practice	5. Consistent with other studies
(Anderson *et al.* 2002)	no	yes	yes	possibly	yes
(Beuche *et al.* 2007)	no	partially	yes	no	yes
(Doernhoefer 2006)	Admits some bias	Partially (expert opinion)	yes	yes	yes
(Eriksson *et al.* 2005)	no	yes	yes	possibly	yes
(Grimm 2003) **	no	no	no	Not clear	Not clear
(Heumesser and Houdek 2003)	no	yes	yes	possibly	yes
(Woit 2005)	no	unclear	possibly	possibly	Not clear
(Dick 2005)**	yes	Expert opinion	No	No	unclear
(Kealey *et al.* 2006)	no	Yes	yes	yes	yes
(Stevens 2001)**	yes	Expert opinion	no	no	unclear

4.4 Further advice

1. Further advice on step 3 is available in the *IEEE Software* article [1] on Evidence Based Software Engineering.

2. Keep your appraisal focused on the EBSE question, particularly on the outcome of interest that you have identified in your EBSE question.

3. For relevance, consider the degree to which the article(s) being appraised refer in particular to the outcome of interest, to the intervention and to the comparison/baseline.

4. Overall, you are being asked to judge the quality of each information resource individually, and *then* to make a tentative recommendation about whether to adopt or not adopt the tool or tools you are evaluating. Your tentative recommendation should be based on what the information resources conclude *overall* about the outcome of the tool(s).

5. Different types of article will need to be evaluated in different ways. For example, a blog article is unlikely to present empirical evidence in the same way that a research-based article presents empirical evidence. Similarly, the arguments made in a blog entry may be more relevant but may be less reliable.

6. The checklist provided on p. 62 of the *IEEE Software* article [1] is most appropriate for published research and it may not apply (or only parts of it may apply) to other kinds of information. Table 6 presents a generic checklist that is intended to complement the IEEE Software article checklist. Table 5 presents an example checklist that is specific to the requirements management tool example used earlier in these Guidelines.

7. Consider using the guidelines of systematic reviews to help you with this step of your evaluation. (An article on systematic reviews is referenced in the General Advice section of

these guidelines.) You will not have the time however to undertake a full-scale systematic review.

8. One way to integrate evidence is to perform a simple vote count e.g. count the number of articles that conclude that the tool does improve the outcome of interest, and the number of articles that conclude that the tool does not improve the outcome of interest; then find which count is highest. You could also count other things e.g. the number of articles that conclude that the comparison/baseline technology improves the outcome, or the number of articles that do not find a clear conclusion.

9. Remember that the items that you appraise in step 3 do not just have to be published, research articles. An article may be a blog entry, an email, the results of questionnaire, a web page etc.

Table 5 Another example checklist

	Question	Values[1]
1	When was the resource published?	Date
2	Where is the resource published?	Academic journal; Academic conference ; Thesis [MSc, PhD]; Technical report / Working paper; Trade journal; White paper; Other report; Online webpage
3	What is/are the affiliation(s) of the author(s)?	Academia: faculty; Academia: student; Vendor/producer; Consultant; Government research centre (e.g. Defence, SEI); Industry 'user'/consumer
4	What is the primary purpose of the resource?	To report the tool evaluation(s); to report on the development of an evaluation methodology; other
5	To what degree is the evaluation methodology described in the resource?	No description Some description Distinct description e.g. as a methodology section
6	What is the focus of the analysis in the evaluation?	On a cause-effect relationship; on the features/attributes of the tool; Other focus

[1] For conciseness, the allowable values are presented as a list, with each value separated by a semi-colon. In some cases, sub-values are also 'allowed' and these are indicated in square brackets i.e. []. Examples of acceptable values are also sometimes included for clarification and these are presented in round brackets i.e. ().

Table 6 An example checklist for appraising resources relating to a specific EBSE question

#	Available answers (select one from those available)
0	Paper reference
1	What year was the paper published?
2	Is the paper written by people in academia, industry, or academia and industry?
3	Is the paper written by Borland?
4	How many RMTs[2] does the paper refer to?
5	Does the paper specifically consider Borland's CaliberRM
6	Does the paper provide a comparison of two or more RMTs?
7	Does the paper *analyse* the features of one or more RMTs?
8	Does the paper specifically *analyse* the Traceability feature?
9	Does the paper *analyse* the effect, or impact, of the traceability feature?
10	If your answer to Q9 in this table was "Yes", what is the effect or impact being investigated?
11	To what degree is an evaluation method for evaluating the RMT(s) described in the paper?
12	What is the primary purpose of the paper, as stated by the paper?

5 Guidelines for EBSE Step 4

5.1 Brief description of the step

Integrate the findings from the appraisal in step 3 with practical experience (your own, or others) and with any particular values and circumstances specific to the task situation.

5.2 Guidelines

1. State any activities, projects, or jobs (paid or unpaid) that you have previously undertaken and that relate to evaluations and to software engineering in some way. For example, a placement year, coursework assignments, personal development of software, software projects, technology assessments.

2. State activities that you have undertaken that relate to:

   - Tool use in general
   - Tool selection or evaluation in general
   - The specific tool(s) identified in the EBSE question
   - The specific problem being considered here e.g. the requirements process
   - The outcome of interest defined in the EBSE question

[2] RMT = Requirements Management Tool

3. Compare your personal experiences with the expertise and experience you have defined for the technology user in the EBSE question.

4. From the lists generated by guidelines 1 and 2 above and your comparison in guideline 3 identify and state those activities that are relevant to this evaluation and explain how they are relevant.

5. Consider any specific values or circumstances that may be particularly relevant for the current evaluation, and state those.

6. Combine the information you have gathered through guidelines 1 – 5 above with the tentative recommendation from EBSE step 3.

7. Make a recommendation on whether one should or should not adopt the tool(s) you have evaluated. Provide a brief argument to support your recommendation. The recommendation and argument should be based on the results of your evaluation in EBSE step 3 and/or the results of your integration here in EBSE step 4.

5.3 Example(s)

"During my placement, I had experience of gathering and recording requirements as part of developing a large-scale accountancy system. From this experience I learned [a particular lesson] about the importance of [some aspect] of [a tool or process]. This lesson agrees with the evidence provided in [article 1] and [article 2] that were appraised in step 3. My personal experience therefore is consistent with the appraised evidence and I continue to recommend that [a tool] is used to improve [the outcome]."

5.4 Further advice

1. Further advice on step 4 is available in the *IEEE Software* article [1] on Evidence Based Software Engineering.

2. Practitioners, students and researchers can find it hard to reflect on the activities they have undertaken in the past and how their experience from those activities is relevant to an EBSE evaluation. Therefore, ensure you spend a suitable amount of time and effort identifying and reflecting on your activities.

3. The application of guidelines 1 – 3 of EBSE step 4 can help you demonstrate that you have experience, even if you then explain through the application of guidelines 4 – 6 that this experience is not relevant. It is important that you demonstrate that you have considered your practical experience as part of the overall evaluation.

6 Guidelines for EBSE Step 5

6.1 Brief description of the step

Review your use of the EBSE guidelines.

6.2 Guidelines

1. There is always a possibility that the recommendation made by the evaluator turns out, in time, to be incorrect. After all, the evaluator is being asked to make a recommendation at a particular point in time on the basis of evidence and experience available to them at that time. The purpose, therefore, of EBSE step 5 is to encourage the evaluator to reflect on the evaluation they have conducted so that they can improve both their performance during future evaluations and the outcome of (or: the recommendation that results from) those evaluations. As an evaluator you should reflect on the following:

    - The evaluation guidelines themselves
    - Your interpretation and use of the guidelines
    - The task situation, and how you converted that task situation into an EBSE question
    - The time and effort you planned for each step of EBSE, the time and effort actually expended on each step, and explanations for the differences between the plan and the actual. (See General Advice for further information.)

2. To assist in your evaluation, the EBSE guidelines [1] recommend two types of review: an After Action Review (AAR) and a Post-Mortem Analysis (PMA). Consult the *IEEE Software* article for more information on these two types of review.

3. To help with your step 5 evaluation, you could also consider the following questions:

    - How did the supplementary guidelines provided here make it easier or harder to undertake an EBSE evaluation?
    - When would you use EBSE again?
    - When would you not use EBSE again?
    - How could the EBSE supplementary guidelines be improved, and why?
    - Whether (and why) EBSE would benefit from using materials (e.g. guidelines, templates, checklists) of other evaluation methodologies.

6.3 Example(s)

The following example was taken from an evaluation where the evaluator used the EBSE Supplementary Guidelines to evaluate the Telelogic DOORS Requirements Management Tool (RMT). The Evaluator then used the AAR and PMA to reflect on the evaluation:

> "See Appendix ... for a full reflection of what the methodology has taught me. In this section I
>
> follow the guidelines given in (Dybå *et al.* 2005) to evaluate my performance.
>
> 1. After Action Review
>
> 1.1 What was supposed to happen?
>
> I hoped to gather and analyse peer-reviewed work in a systematic way that evaluated a specific RMT against certain features in fit the needs of our customer. I was going to integrate this evidence with my own knowledge and experience of RMTs to give a recommendation to the customer.
>
> 1.2 What actually happened?
>
> I could not source many evaluations of the tool and was unable to give a clear recommendation
>
> 1.3 Why were there differences?
>
> Researchers are more likely to publish work on new tools or by tweaking existing tools. Evaluation exercises are not often performed and published in software engineering
>
> 1.4 What did we learn?
>
> That there is a gap in the type of publications that practitioners are likely to find very useful
>
> 2. Post Mortem Analysis
>
> 2.1 What went so well that we want to repeat it?
>
> Following the guidelines taught me a systematic way of gathering secondary data, also defining a research question and its component parts gave an important focus to the work on which to build
>
> 2.2 What was useful but could have gone better?
>
> The ACM database is a very useful resource, but would be better if we could have nested and or Boolean searches. Also if the ACM database could provide references that could be directly imported into Endnote would save lots of time (had to be done manually – which is very time consuming). Perhaps using Google Scholar might have been a better resource –although the selection criteria might be difficult as we are likely to get too many hits.
>
> 2.3 What were the mistakes that we want to avoid for the future?

Perhaps starting with a more general question (top down) and then narrowing the scope would be a quicker way to find relevant publications in the area, rather than define a precise question (bottom up) that resulted in no papers returned. Also considering less rigorous form of resource (such a blogs) may have resulted in more relevant material for our needs.

2.4 What were the reasons for the successes or mistakes, and what can we do about them?
2.4.1

What were the reasons for the success or mistakes? The process is rigorous and the material gathered is reliable which I consider a success, however it is perhaps a mistake to use this resource when there was little 'evaluation' of the tool in question to draw on

2.4.2 What can we do about them?

Try to explore each resource quickly (e.g. blogs, e-journals) along with the more time consuming searches for the peer-reviewed literature to establish which resource best suits our needs.

Although I didn't have direct experience of using the RMT the guidelines allowed me to integrate my personal opinion about tool adoption with the external evidence. Following EBSE methodology is time consuming, and although I only used a subset of the customer's requirements, feel that it is a good basis on which to begin to make a recommendation."

6.4 Further advice

1. Further advice on step 5 is available in the *IEEE Software* article [1] on Evidence Based Software Engineering.

2. Practitioners, students and researchers often find it hard to reflect on their use of EBSE. The questions presented above are designed to help you reflect on your use of EBSE.

Evidence-based evaluation of the Supplementary Guidelines

7 Introduction

This section reports results from an evaluation of students' use of the first version of the EBSE Supplementary Guidelines. The evidence presented here provides further guidance on how to conduct evaluations.

8 Brief explanation of the evaluations

The Supplementary Guidelines were first used by the 2007 cohort of students on the *Empirical Evaluation in Software Engineering* module at the University of Hertfordshire, as part of a coursework on EBSE. During a tutorial after students had submitted their EBSE courseworks, the first author of the supplementary guidelines sought feedback from the students on the coursework and the guidelines. Students were invited, but not required, to complete a two-page feedback questionnaire and most but not all students at the tutorial did so. In total 12 completed questionnaires were returned.

9 Evaluator recommendations resulting from an evaluation

Table 7 What recommendation did the evaluators (students) make? (analysis of the 37 courseworks from the 2007 cohort)

Recommendation	Count	Percentage
Intervention	23	62%
Baseline	5	14%
Ambiguous	6	16%
No recommendation	3	8%
Total	37	

Table 7 indicates that most of the students ultimately recommended the intervention that they had chosen for their EBSE question (the structure of an EBSE question is described earlier in the supplementary guidelines). There is a well-established psychological bias in human thinking, called confirmation bias, where an evaluator seeks evidence to *support* an already-adopted claim or position. These students appear to be exhibiting confirmation bias. Therefore, care should be taken by an evaluator to seek both supporting *and* contradictory evidence on both the intervention and the baseline being considered in the evaluation.

All 12 students of the 2007 cohort who completed the feedback questionnaire responded that they made a recommendation in their coursework. Verbal feedback provided during the tutorial, but not recorded on the questionnaire, suggests that at least some students made a recommendation

because the Supplementary Guidelines required it rather than the student being confident about the making of a recommendation. This finding suggests that an investigator of EBSE should be careful about attributing students' recommendations only to confirmation bias. Again, an EBSE evaluator should take care to either make a recommendation that is based on the evidence evaluated, or alternatively to state that no recommendation can be made on the basis of the evidence.

Three recommendations arise from the evidence presented above:

1. Be aware of the potential threat of confirmation bias
2. Seek evidence to both support an undermine/contradict the proposed intervention
3. Do not make a recommendation just because the Guidelines say that you should

10 The degree of challenge of the coursework

At the tutorial for the 2007 cohort of students, we asked the students to indicate, on a scale of 1 – 7, how challenging they found the coursework. A value of 1 represented the statement "The easiest coursework I have ever done" and a value of 7 represented the statement "The hardest coursework I have ever done". For the sample of 12 students, the median and mode was 6 and the mean was 6.25. The lowest value was 5 and one student responded by extending the scale to 8! Re-setting that value to 7, from 8, actually changes the modal value from 6 to 7! and very slightly reduces the mean value to 6.16. Overall, this sample of students is clearly indicating that this was the hardest coursework that they have had to undertake on their degree programme. At the same time, this is the final year of the degree programme and this coursework was the first coursework that the students had been set in their final year. It may be that students would revise their opinion of the coursework having completed other courseworks, including the final year dissertation.

The evidence on the degree of challenge of the coursework indicates that novice users of EBSE should expect an evaluation to be very demanding.

11 The easiest and hardest steps in EBSE

At the tutorial for the 2007 cohort of students, we asked students what they considered to be the easiest and hardest stages in EBSE. Table 8 presents the results. The table indicates that students could more clearly identify *one* easiest step but could not clearly identify only one hardest step. Because the students had difficulty identifying only one hardest step, Table 8 reports two columns for the hardest step together with a total. The steps considered to be easiest were Step 4 and Step 1. These were the opposite of the hardest stages i.e. Step 2 and Step 3.

Table 8 The easiest and hardest steps in EBSE

EBSE step	Easiest choice 1	Hardest choice 1	choice #2	Total
Step 1	4	0	0	0
Step 2		7	0	7
Step 3		4	4	8
Step 4	6	0	1	1
Step 5	2	1	0	1

The evidence presented in the table suggests that an evaluator should prepare carefully for the evaluation and should ensure that they allocate a reasonable proportion of the resources available for the evaluation to Steps 2 and 3.

12 The usefulness of the resources

At the tutorial for the 2007 cohort of students , we also asked students whether they received sufficient resources to help them with their coursework. We used a scale of 1 – 3, with a value of 1 indicating 'Not enough support', a value of 2 indicating 'Enough support', and a value of 3 indicating 'More than enough support'.

Table 9 The adequacy of resources

Resource	Median	Mode
The lectures and tutorials on EBSE	2.5	3
The Supplementary Guidelines on EBSE	2.6	3
The EBSE articles posted on StudyNet	2.3	3
Other articles posted on StudyNet	2.3	2
StudyNet Classroom Discussion	2.1	2
Personal emails with tutor	2.1	2

Note: Responses were on a 3-item scale, where:
1= Not enough support; 2 = Enough support; 3 = More than enough support

Table 9 indicates that, generally, students thought the resources were sufficient, with the Supplementary Guidelines receiving the highest 'score'. The table provides evidence that the supplementary guidelines were beneficial to the students in undertaking their EBSE evaluations.

The Assessment Scheme

13 Introduction

We do not report any criteria here for assessing the importance of each element of the scheme. Related to the lack of criteria, we do not report a specific assessment mark allocation here for each element of the scheme, which may for example be used to assess coursework. The allocation of marks to each element in the scheme is a decision that we believe that individuals should make based on their use of the scheme. For example, a lecturer may determine a particular mark allocation based on the coursework that they set (and may choose not to reveal that to the students until the assessment is complete). Conversely, a researcher may allocate marks based on the particular aspects of the EBSE methodology that they want to empirically investigate. Finally, a practising software professional may use the assessment scheme to self-check the evaluation being undertaken.

14 The elements of the Assessment Scheme

14.1 EBSE Step 1

Elements of the coursework to award marks

Stating an EBSE question comprising five sensible components. Students must provide an EBSE question and the 5 marks are only available for that question.

Appropriate definitions. Students should provide a definition of outcome, but may also provide other definitions.

The following mark allocation is indicative

Explanations, based on problem(s) identified

Explanation of choice of intervention

Explanation of choice of outcome

Assumptions

Statement of pre-conceived expectation of performance that specifically relates to the outcome of interest

14.2 EBSE Step 2

Elements of the coursework to award marks

Identification of information source available, accessible and feasible to use
Statement of information sources to be used in search, together with explanation
Statement of search strings, to include:
- Keyword(s) in one search
- Multiple search strings

Statement of inclusion and exclusion criteria
- e.g. date range for inclusion/exclusion
- e.g. including/excluding blogs

Summary of searches including search results, to include:
- Search engine e.g. Google, Scholar, Blog, Voyager
- Keywords
- Date of search
- Number of results

Sensible refinement and repetition of searches (where this indicates that the student is improving their searches i.e. narrowing the results)
List of information selected for appraisal and explanation of their selection
A clear focus on identifying information that is rigorous and relevant
A range of information sources searched e.g. journal articles and blogs and user forums and trade magazine articles etc.

14.3 EBSE Step 3

Elements of the coursework to award marks

The resources identified in step 2 are appraised in a balanced way i.e. step 3 does not just report on the appraisal of a subset of the articles, but provides a balanced appraisal of all of them
Individual articles are clearly appraised (prior to integration of evidence and arguments)
Articles are appraised for their (methodological) rigour e.g. examining vested interests, bias of data etc.
Articles are appraised for their relevance to the five components of the EBSE question
Appropriate use(s) of the EBSE checklist
Statement(s) of retention or rejection of article
Integration of results from all articles e.g. through vote-counting; through argument; through meta-analysis!; through detailed comparison of results from individual articles.
Statement of tentative recommendation

14.4 EBSE Step 4

Elements of the coursework to award marks

Statement of activities etc.
Statement of activities
Comparison of experiences. There must be a comparison!
Statement of activities relevant to evaluation
Consideration of specific values and circumstances, with regards to (for example):
- The student's values and circumstances
- The company's values and circumstances
- Values and circumstances raised from articles in step 3

Recommendation
Argument to support recommendation: internal quality of argument
Argument to support recommendation: consistency with evidence in steps 2 and 3

14.5 EBSE Step 5

Elements of the coursework to award marks

Evaluation of the guidelines themselves
Evaluation of the application of the guidelines
The problem
The planned and actual time and effort, and explanation of difference
The remaining 8 marks are available for consideration of the following:
Application of AAR, PMA, suggested questions or appropriate alternatives
The following questions (or similar)
- Did the supplementary guidelines provided here make it easier or harder to undertake an EBSE evaluation? Explain why.
- When would you use EBSE again?
- When would you not use EBSE again?
- How could the EBSE guidelines be improved, and why?
- Whether (and why) EBSE would benefit from using elements (e.g. guidelines, templates, checklists) of another evaluation methodology, such as DESMET.

References

1. Dybå, T., B.A. Kitchenham, and M. Jorgensen, Evidence-based software engineering for practitioners. *IEEE Software*, 2005. **22**(1): p. 58-65.
2. Kitchenham, B., Procedures for Performing Systematic Reviews, Keele University Technical Report TR/SE-0401 and NICTA Technical Report 0400011T.1, 2004, Keele University (UK) and National ICT Australia Ltd. p. 33.
3. Kitchenham, B. and S. Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering (Version 2.3), EBSE Technical Report EBSE-2007-01, 2007, Keele University and Durham University.
4. Rainer, A. and S. Beecham, A comparison of three literature searches for evaluations of requirements management tools, Technical Report UH-CS-TR-480, 2008, University of Hertfordshire: Hatfield, U.K.
5. Rainer, A. and S. Beecham. A follow-up empirical evaluation of evidence based software engineering by undergraduate students. in *12th International Conference on Evaluation and Assessment in Software Engineering*. 2008. University of Bari, Italy, 26-27 June 2008.
6. Rainer, A. and S. Beecham, Supplementary Guidelines and Assessment Scheme for the use of Evidence Based Software Engineering, Technical Report CS-TR-469, 2008, University of Hertfordshire: Hatfield, U.K.
7. Rainer, A., S. Beecham, and T. Hall, Assessing undergraduate students' use of Evidence Based Software Engineering, CS-TR-462, 2007, University of Hertfordshire: Hatfield, UK.
8. Rainer, A., S. Beecham, and T. Hall. Assessing undergraduate students' use of Evidence Based Software Engineering: a preliminary evaluation of a research instrument. in *Workshop Series on Empirical Software Engineering (WSESE'07)*. 2007. Riga, Latvia, 2nd - 4th July 2007.
9. Rainer, A., S. Beecham, and C. Sanderson. An assessment of published evaluations of requirements management tools. in *Evaluation and Assessment in Software Engineering (EASE)*. 2009. Durham, UK, 20-21 April 2009.
10. Rainer, A., D. Jagielska, and T. Hall. Software engineering practice and evidence based software engineering research. in *International Conference on Software Engineering (ICSE) Workshop on Realising Evidence Based Software Engineering (REBSE)*. 2005. St Louis, Missouri, USA, 17 May 2005.
11. Rainer, A.W., T. Hall, and N. Baddoo. A preliminary empirical investigation of the use of evidence based software engineering by under-graduate students. in *2006 Evaluation and Assessment in Software Engineering (EASE) Conference,*. 2006. Keele University, U.K., 10th – 11th April, 2006.